Free3D: Consistent Novel View Synthesis without 3D Representation

Chuanxia Zheng Andrea Vedaldi
Visual Geometry Group, University of Oxford
{cxzheng, vedaldi}@robots.ox.ac.uk

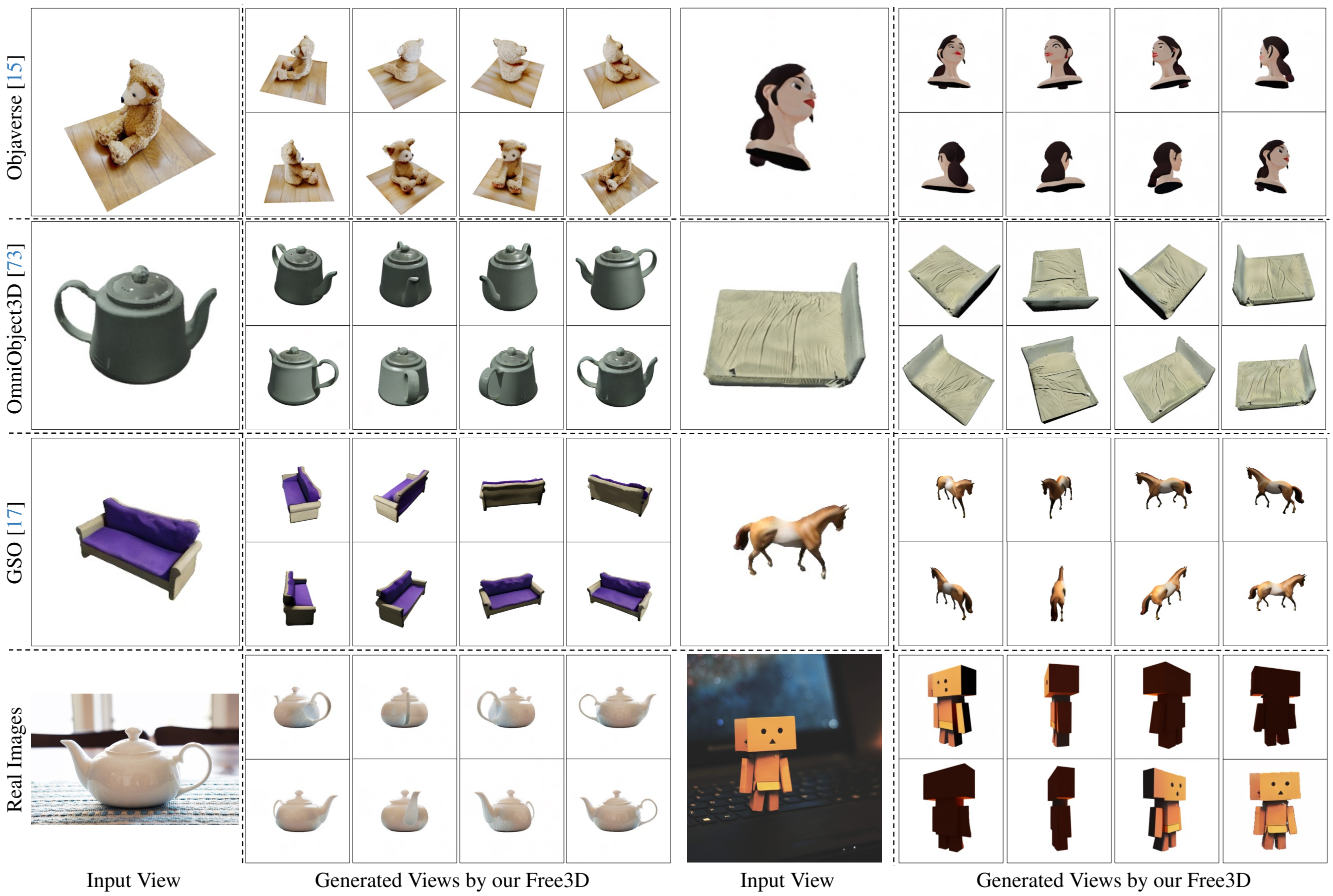


Figure 1. Given a single input view, **Free3D** synthesizes consistent 360° views accurately without using an explicit 3D representation. Trained on `Objaverse` only, it generalizes well to new datasets and categories.

Abstract

We introduce Free3D, a simple accurate method for monocular open-set novel view synthesis (NVS). Similar to Zero-1-to-3, we start from a pre-trained 2D image generator for generalization, and fine-tune it for NVS. Compared to other works that took a similar approach, we obtain significant improvements without resorting to an explicit 3D representation, which is slow and memory-consuming, and without training an additional network for 3D reconstruction. Our key contribution is to improve the way the target camera pose is encoded in the network, which we do by introducing a new ray conditioning normalization (RCN) layer. The latter injects pose information in the underlying 2D image generator by telling each pixel its viewing direction. We further improve multi-view consistency by using light-weight multi-view attention layers and by sharing generation noise between the different views. We train Free3D on the Objaverse dataset and demonstrate excellent generalization to new categories in new datasets, including OmniObject3D and GSO. The project page is available at https://chuanxiaz.com/free3d/.

1. Introduction

Novel view synthesis (NVS) has seen significant recent progress [4, 5, 9, 33, 44], in part due to new neural representations like NeRF [43]. However, many such NVS methods require to optimize a 3D model from scratch *for each scene*, and require dozens of input views to work well; they are thus impractical in many applications. This has motivated authors to apply generative models [20, 34, 61] to the NVS task, eschewing the need for an explicit 3D model and enabling NVS from a single image [27, 29, 42, 45, 63, 68, 78]. Most of these learn 3D priors that are applicable to an entire *object category*, or even to *unstructured collections of objects* [27, 29, 42, 45, 63, 68, 78]. In this work, we aim at improving the quality of these approaches while also generalizing further their applicability by considering an *open-set* setting, where at test time one is given not only new object instances and categories, but also new datasets.

There are two quality targets for NVS: **(i)** The output must accurately reflect the pose of the target cameras, and **(ii)**, when several views of the same object are generated, they must be mutually consistent. In order to achieve these goals, recent methods [37, 39, 57, 75, 76] build a 3D representation of the object or scene, often combined with a pre-trained 2D generative model [51]. Using a 2D model with a 3D representation works well but adds complexity.

In this paper, we introduce Free3D, a simple and efficient method that can also achieve consistent NVS results *without the need to rely on an explicit 3D representation*. Zero-1-to-3 [38] and its precursor 3DiM [68] are perhaps the best-known examples of such a 3D-free NVS system. Like Zero-1-to-3, Free3D builds upon a pre-trained 2D generative model like Stable Diffusion [51], trained on a large-scale dataset (*i.e.* LAION 5B [56]), as a data prior. The prior knowledge contained in such a 2D generator is extremely important to be able to 'guess' plausible novel views of open-set objects, which is inherently highly ambiguous. However, we show empirically that Zero-1-to-3 has, in practice, poor camera pose control, and, when tasked with generating multiple views, not very consistent. The latter is unavoidable in their design because each view is sampled independently from scratch and, owing to the ambiguity of reconstruction, there is no reason why compatible images would be generated each time.

To mitigate these issues, we first show that better camera control can be achieved by switching to a different representation of camera pose. Specifically, we introduce a *ray conditioning normalization* (RCN) layer (Fig. 2 (b)) which tells each pixel its viewing direction. This is a distributed representation of the camera, which should be contrasted to the concentrated camera representation used in [38], where the camera pose is processed as language-like tokens that may be difficult for the network to interpret and utilize [55]. In contrast, with our RCN layer, we show how to effectively incorporate this per-pixel ray information in an *existing* text-to-image diffusion model, which empirically leads to significantly more accurate NVS in our experiments (Tabs. 1 and 2). RCN is inspired by the design of methods like NeRF [43], LFNs [60] and 3DiM [68], which also work by modelling individual rays.

While RCN allows to control the camera more accurately, it does not improve multi-view consistency. For the latter, we introduce a *pseudo-3D cross-view attention* module (Fig. 2 (c)) which is inspired by video diffusion models [6, 7, 21, 25, 58, 72]. This layer fuses information across all views instead of processing them independently. Furthermore, we use *multi-view noise sharing* when generating the different views of the object, which further enhances consistency due to the continuity of the denoising function, reducing aleatoric variations between the views.

We benchmark Free3D against state-of-the-art methods [14, 38, 39, 69] on *open-set* NVS. Although the model is trained on only one dataset, it generalizes well to all recent NVS benchmark datasets, including `Objaverse` [15], `OmniObject3D` [73], and Google Scanned Object (`GSO`) [17]. A thorough experimental assessment shows that our approach consistently outperforms existing methods, both quantitatively and qualitatively.

To summarise, with Free3D we make the following contributions: (i) We introduce the *ray conditioning normalization* (RCN) layer and show that representing the camera by utilizing a combination of distributed ray conditioning and concentrated pose tokens significantly improves pose accuracy in NVS. (ii) We show that a small *multi-view attention* module is sufficient to improve multi-view consistency by exchanging information between views at a low cost. (iii) We find that *multi-view noise sharing* further improves consistency. (iv) We demonstrate empirically that Free3D achieves consistent NVS *without needing a 3D representation* and outperforms the existing state-of-the-art models on both pose accuracy and view consistency.

2. Related Work

Per-Scene NVS. Early NVS works relied on epipolar geometry to interpolate between different views of the same scene [11, 13]. A recent breakthrough was to represent 3D scenes as implicit neural fields, as proposed by SRN [59], DeepSDF [46], NeRF [43] and LFN [60], and further improved in follow-ups [4, 5, 9, 19, 33, 44, 79]. Even so, data efficiency, generalizability, and robustness remain a limitation: such systems require multiple views to learn the 3D representation from scratch for every scene.

To bypass the need for multiple input views, DreamFusion [47] proposes to distill 3D models from a large-scale pre-trained 2D diffusion model [52]. RealFusion [41] extends the latter to single-view image reconstruction by adding the input image as a constraint during distillation.

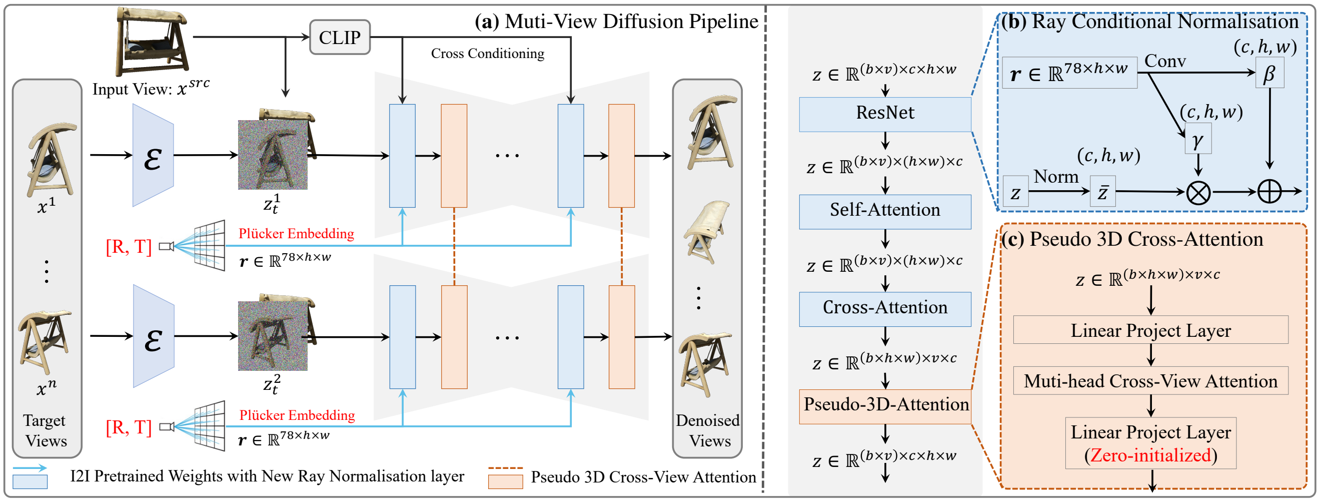


Figure 2. **The overall pipeline of our Free3D.** (a) Given a single input image, the Free3D jointly predicts multiple target views, instead of processing them independently. (b) We propose a novel *ray conditional normalization* (RCN) layer, which uses a *per-pixel* oriented camera ray to module the latent features, enabling the model's ability to capture more precise viewpoints. (c) A memory-friendly *pseudo-3D cross-attention* module is introduced to efficiently bridge information across multiple generated views.

Several follow-up works [48, 49, 55, 64] further improve the resolution and quality of the resulting 3D assets. Although these models are open-ended [41], they still require lengthy per-scene optimisation.

Category-Specific NVS. Some recent work [12, 45, 65, 70] built the autoencoder architecture for NVS. Driven by the effectiveness of light field rendering [1, 2], other NVS approaches [53, 54, 62] query a network for colour of different rays. 3DiM [68] then introduced diffusion models into NVS, an approach also taken by RenderDiffusion [3], HoloDiffusion [31], SparseFusion [82], GeNVS [8], Viewset Diffusion [63] and LFD [74]. However, these methods train their models from scratch using pure 3D data, which are too small to afford open-set generalization, and are thus limited to one or few categories.

Open-set NVS. To operate in an open-set setting, Zero-1-to-3 [38] builds on a large pre-trained 2D image generator and fine-tune it on the Objaverse dataset [15]. While it has good generalizability, it fails to achieve high pose accuracy, and its reconstructions are inconsistent across views. To mitigate these issues, other concurrent approaches either integrate 3D representations into the network [30, 39, 76] or train another 3D network [36, 37, 57, 69, 75]. These methods can output high-quality target views, but are computationally expensive. More importantly, these methods do *not* address the issue of pose representation, which we find to be a key bottleneck in NVS. In contrast, our method is *3D-free* and achieves comparable or better NVS quality, due to *ray conditioning normalisation*, *multi-view attention*, and *multi-view noise sharing*. A more recent concurrent work is SVD [6], which synthesizes multi-view videos, but is trained with fixed camera poses. In contrast, Free3D is trained to generate arbitrary 360° viewpoints. The follow-up work SV3D [67] builds on a similar concept.

3. Method

Our goal is to learn a model Φ that, given as input an image x^{src} and a sequence of camera poses $\mathcal{P} = \{\mathbf{P}^i\}_{i=1}^N$, synthesizes corresponding novel views $\{x^i\}_{i=1}^N$ which are accurate and consistent without relying on an explicit 3D representation. We generate all the views together, conditioned on the given inputs, so that the network has a chance to produce several consistent views together.

Specifically, we address two important challenges: **(i)** ensuring that the model accurately captures the pose of the target view and **(ii)** ensuring the different views consistent in terms of geometry and appearance. To achieve these, our framework, illustrated in Fig. 2, extends a 2D generator by injecting at each layer a ray-conditional normalization layer (Sec. 3.1) and a pseudo-3D attention layer (Sec. 3.2). It also utilizes multi-view noise sharing (Sec. 3.2). The former captures pose more accurately, whereas the last two improve multi-view consistency. Although 3DiM [68] also passes ray information to the network, the key difference is that we show how to inject ray conditioning in a *pre-trained* 2D generator. The benefit is to marry generalization with high pose accuracy without an explicit 3D representation or an additional 3D reconstruction network.

3.1. Ray Conditioning Normalization (RCN)

Ray Conditioning Embedding. Given a target view $\mathbf{P} = (\mathbf{K}, \mathbf{R}, \mathbf{T})$, for each pixel (u, v) in the image, we define the

Plücker coordinates $\boldsymbol{r}_{uv} = \phi(\boldsymbol{o}, \boldsymbol{d}_{uv}) = (\boldsymbol{o} \times \boldsymbol{d}_{uv}, \boldsymbol{d}_{uv}) \in \mathbb{R}^6$ of the ray going from the camera center $\boldsymbol{o} \in \mathbb{R}^3$ through the pixel, where $\boldsymbol{d}_{uv} = \mathbf{R}^\top(\mathbf{K}^{-1}(u, v, 1)^\top - \mathbf{T}) \in \mathbb{R}^3$ is the ray direction and $(\mathbf{K}, \mathbf{R}, \mathbf{T})$ are the camera's calibration, rotation and translation parameters. This encoding was originally introduced by LFN [60]. It is invariant to shifting the camera along the ray, meaning that $\phi(\boldsymbol{o} + \lambda \boldsymbol{d}, \boldsymbol{d}) = ((\boldsymbol{o} + \lambda \boldsymbol{d}) \times \boldsymbol{d}_{uv}, \boldsymbol{d}_{uv}) = (\boldsymbol{o} \times \boldsymbol{d}_{uv}, \boldsymbol{d}_{uv}) = \phi(\boldsymbol{o}, \boldsymbol{d})$, which matches the fact that light propagates in straight lines.

Ray Conditioning Architectures. Our goal is to modify the denoising neural network $\hat{\epsilon}(z, t, y)$ so that the conditioning information y includes ray conditioning. We experiment with a number of different architectures to do so, proposing various ray conditioning layers:

- *Concatenation.* Following Zero-1-to-3 [38], a natural choice is to concatenate the noised target z_t^{tgt}, original source embedding z^{src}, and ray conditioning at the input, which is then $(z_t^{\text{tgt}}, z^{\text{src}}, \boldsymbol{r})$ instead of z_t^{tgt} alone.
- *Multi-scales concatenation.* We further consider concatenating the ray embeddings $\boldsymbol{r}$ to each intermediate layer in the UNet ϵ_θ. Note that each layer operates at a different resolution, so this amounts to injecting the information $\boldsymbol{r}$ at different scales.
- *Ray conditioning normalization (RCN).* Finally, we propose to combine the adaptive layer norm [18, 26, 32, 81] with ray conditioning to modulate the image latents. Specifically, the activation latent F_i of the i-th layer in the UNet ϵ_θ is modulated by:

$$\text{ModLN}_{\boldsymbol{r}}(F_i) = \text{LN}(F_i) \cdot (1 + \gamma) + \beta, \tag{1}$$

  where $(\gamma, \beta) = \text{MLP}^{\text{mod}}(\boldsymbol{r})$ are *scale* and *shift* parameters predicted from the ray embeddings $\boldsymbol{r}$. This is applied to each sub-module of the UNet ϵ_θ (Fig. 2 (a) & (b)).

Interestingly, while RCN works best, we show empirically that all such architectures lead to strong improvements (Tab. 3), which confirms that the key factor in better camera control is to inject ray information in the network.

Discussion. RCN differs significantly from Zero-1-to-3 [38] and follow-ups [28, 37, 39, 67, 69, 76] which encode the camera pose $\mathbf{P}$ as *global* tokens. Ray-Cond [10] and LFD [74] also applied ray conditioning for NVS in GAN and Diffusion, respectively, but they only concatenate it as additional channels, and consider category-specific NVS. Similar to *ray-casting* [40, 43], RCN considers one ray per pixel, but, differently from the, does *not* evaluate hundreds of samples per ray. Because of this, Free3D dramatically reduces the rendering time and memory consumption compared to concurrent works like [37, 39, 76].

3.2. View Consistent Rendering

Given a source image x^{src}, our goal is to render a series of consistent novel views x^i. While the camera encoding technique of Sec. 3.1 significantly enhances pose accuracy, if images are sampled independently, they will almost never be visually consistent due to the intrinsic ambiguity of the reconstruction task. To remove or at least greatly mitigate this problem, we propose to sample images jointly.

Multi-view attention. We first adapt the frame attention, a well-established method in video generators [6, 7, 21, 25, 58, 72], to capture temporal dependencies across views. As shown in Fig. 2 (c), given a 5D latent $z \in \mathbb{R}^{B \times v \times c \times h \times w}$, we initially reshape it to $z \in \mathbb{R}^{(B \times h \times w) \times v \times c}$, resulting in $batch \times height \times width$ sequences at the length of $views$. Subsequently, this reshaped latent is passed through the pseudo-3D attention module to calculate the similarity across different views. Since this attention layer operates across views but separately for each spatial location, the computational and memory costs are quite low (Tab. 1). Similar to ray conditioning, we inject multi-view attention at each level of the UNet $\hat{\epsilon}_\theta$.

Multi-view noise sharing. In order to reduce the variance between different views, we propose to start sampling each view from the *same* noise vector x_T. The network ϵ_θ still generates different views because it is conditioned on the camera parameters. It can also generate different reconstructions by taking a new noise sample Fig. 7. However, sharing noise reduces the variance *between views*. This can be justified by noting that the network $\hat{\epsilon}_\theta(z_t, t, y)$ is a continuous function of both z_t and y [71]. Besides, different from [66], we do not specially set the noise schedule in different time steps t. More complex designs and training strategies have the potential to improve performance but are not the focus of this work.

3.3. Learning formulation

The learning objective is given by

$$\mathcal{L} = \mathbb{E}_{(Z_0, z^{\text{src}}, \mathbf{P}), \epsilon, t} \left[\|\epsilon - \epsilon_\theta(Z_t, t, y)\|_2^2 \right], \tag{2}$$

where each training sample $(Z_0, z^{\text{src}}, \mathbf{P})$ consist of N encoded target views $Z_0 = \{\mathcal{E}(x^i)\}_{i=1}^N$, the encoded source view $z^{\text{src}} = \mathcal{E}(x^{\text{src}})$ and the viewpoints $\mathbf{P}$. The network is conditioned on the source view and cameras and estimates the noise for all target views jointly. Following [38], we also concatenate the input image code z^{src} to z_t along the channel dimension and add the CLIP [50] encoding.

3.4. Perceptual Path Length Consistency

To quantify the consistency across different views, recent methods [38, 67, 68, 69, 76] trained 3D models from the sampled views and evaluated them on the remaining views. However, this requires training a 3D model for each test instance, which is unfeasible for large-scale testing. Here, we propose to use instead the pairwise perceptual distance [80] to measure the consistency between generated views. In

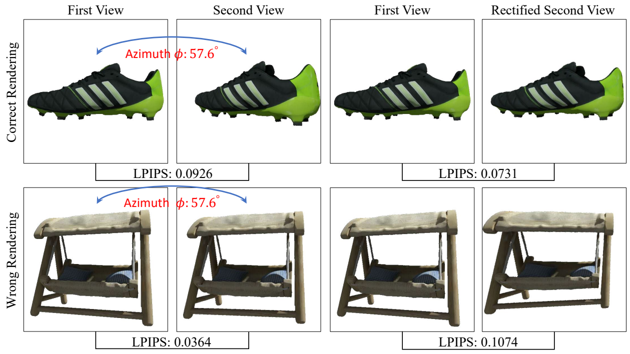


Figure 3. **Perceptual Path Length Consistency (PPLC).** To partly compensate for the viewpoint change, the second image is rectified w.r.t. the first before comparison. To illustrate the importance of using rectification, the figure shows two objects in a large azimuth $\phi : 57.6°$. The top row shows to the left an ideally-rendered image pair, which however attains a large LPIPS score due to the view change. To the right, rectification reduces this score. The bottom row shows the opposite, where a pair of incorrectly rendered views has its LPIPS increased by rectification.

particular, we first subdivide the 360° rendering path into linear segments and then calculate the LPIPS score between two neighboring generated images x^i and x^{i+1}. Naturally, these images differ due to the different viewpoints (Fig. 3). We partly compensate for the viewpoint change by rectifying [22] the second image w.r.t. the first. We then define Perceptual Path Length Consistency (PPLC) of a rendered sequence $\{x^i\}_{i=1}^{N}$ as follows:

$$l_{\text{pplc}} = \mathbb{E}\left[\frac{1}{\phi^2}\|\mathcal{F}(\text{Rect}(x^i)) - \mathcal{F}(\text{Rect}(x^{i+1}))\|_2^2\right], \quad (3)$$

where ϕ is the degree between views x^i and x^{i+1}, which is set as $\phi = 2\pi/50$, with the azimuth 7.2° in all our video rendering. $\mathcal{F}$ is a pre-trained network to ensure the metric matches with human perceptual similarity judgment.

4. Experiments

4.1. Experimental Details

Datasets. For fairness, our model is trained using the exact same protocol as Zero-1-to-3 [38]. They render multiple views for 772,870 objects from `Objaverse` dataset [15]. We use the identical `test` split as they do. To assess how well our mode generalizes to other datasets, and how it compares to other models, we consider two more datasets: `OmniObject3D` [73] and Google Scanned Objects (`GSO`) [17], which contain real-life scanned objects. Since we do not use these datasets for training at all, we use the *entirety* of `OmniObject3D` and `GSO` objects for evaluation (6,000 and 1,030 objects, respectively).

Method	Objaverse[15]				
	SSIM↑	LPIPS↓	FID↓	PPLC↓	Time↓
Zero-1-to-3[38]	0.8462	0.0938	1.52	18.84	**3s/44s**
Zero123-XL[14]	0.8339	0.1098	1.67	25.61	**3s/44s**
SyncDreamer[39]	0.8063	0.1910	7.57	16.32	25s/77s
Consistent123[69]	0.8530	0.0913	1.48	17.89	4s/63s
Ours (Free3D)	**0.8620**	**0.0784**	**1.21**	**10.82**	3s/52s

Table 1. **Comparison with SoTA methods** on all 7,729 objects in `Objaverse test-set`. Recent works like [14, 38, 69] were originally evaluated using a subset of this data only due to the cost of training additional 3D models. Unlike them, we directly evaluate the model on whole `test-set` without using additional 3D networks. The inference time is for rendering a single target view and a 360° video, respectively.

Metrics. We follow [38, 39, 69] and assess the NVS quality by comparing the generated images and the ground-truth views at different levels of granularity, including PSNR, SSIM, LPIPS [80][1], and Fréchet Inception Distance (FID) [23][2]. Some of these metrics look at the statistics of individually reconstructed images rather than exact reconstructions, which is the correct approach given that the NVS task is inherently ambiguous; however, for the same reason, they are *not suitable* for assessing multi-view consistency. Instead, we use PPLC (Sec. 3.4) to evaluate consistency. To do so, we generate 50-frame videos along a pre-defined circular camera trajectory for each object and then compute the PPLC score between neighbouring frames.

Baselines. We compare Free3D to the state-of-the-art Zero-1-to-3 [38] and the follow-ups Zero123-XL [14] SyncDreamer [39] and Consistent123 [69]. While noticed other concurrent works in arXiv for NVS [30, 76, 77], but, as no codes are available during the submission, we do not retrain for comparison.

4.2. Assessing Quality

Quantitative comparison. We first evaluate Free3D and other methods on the `Objaverse` dataset [15, 38], where the model is trained. Unlike previous work [38, 39, 69] which consider only a subset of the test data due to expensive post-processing, we use the *entire test set* of 7,729 3D objects. Quantitative results in Tab. 1 show that Free3D outperforms state-of-the-art models. This includes Zero123-XL [14] and SyncDreamer [39], which are trained on a much larger dataset and with explicit 3D volume representation, respectively. While the concurrent Consistent123 [69] also utilizes a form of multi-view diffusion with cross-view attention, our Free3D significantly improves the quantita-

[1] https://github.com/richzhang/PerceptualSimilarity "*squeeze*"-net.
[2] https://github.com/GaParmar/clean-fid "*clip_vit_b_32*"-net.

(a) Input View **(b)** Target View **(c)** Zero1-to-3 [38] **(d)** Zero123-XL [14] **(e)** SyncDreamer [39] **(f)** Consistent123 [69] **(g)** Ours Free3D

Figure 4. **Qualitative comparisons** on `Objaverse`. Given a target pose, our Free3D significantly improves the accuracy of the generated pose compared to existing state-of-the-art methods. Note that Zero123-XL [14] is trained on the much larger `Objaverse-XL` dataset [14], which contains 10 million 3D objects. More comparisons are provided in the supplement Figs. C.1 and C.2.

(a) Input View **(b)** Target View **(c)** Zero-1-to-3[38] **(d)** Zero123-XL[14] **(e)** SyncDreamer[39] **(f)** Consistent123[69] **(g)** Ours Free3D

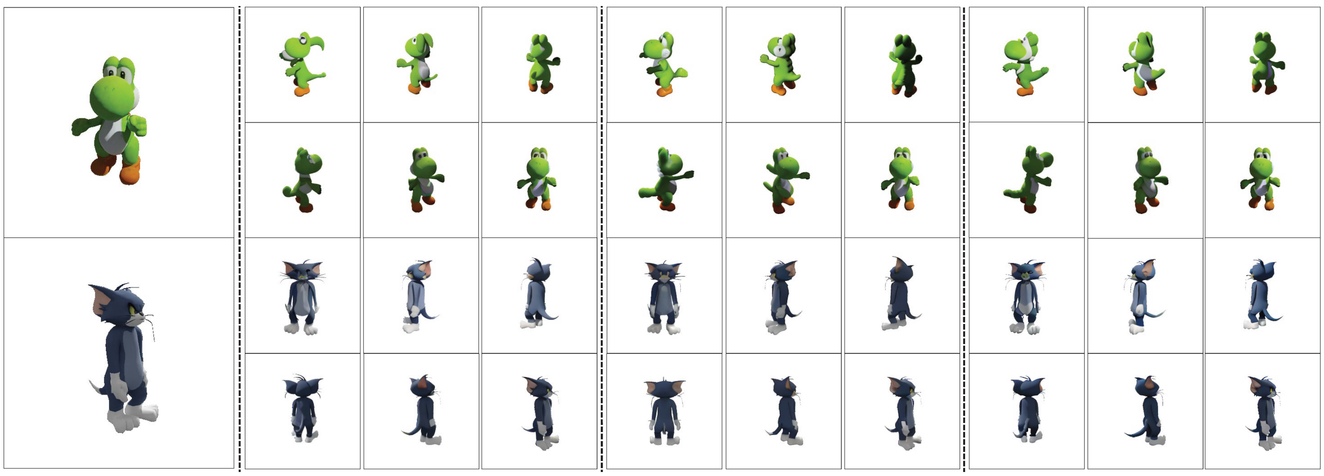
Figure 5. **Qualitative comparisons** on `OmniObject3D` (top two rows) and `GSO` (bottom two rows) dataset. Interestingly, exciting methods cannot deal with unconventional objects, such as the "pie" in the first row, while our Free3D is still robust for such a challenging scenario. More comparisons are provided in supplemental Figs. C.3 and C.4.

tive results (16% relative improvement on LPIPS) on the large evaluation dataset. This suggests that our RCN layer is the primary reason for the observed improvements.

Qualitative comparison. Qualitative results are visualized in Fig. 4. Free3D achieves better results even under challenging viewpoints. SyncDreamer [39] uses an *explicit volumetric representation* representation, but can only generate views with fixed elevation (30°), leading to worse results on synthesizing arbitrary target views. While Consistent123 [69] aims to improve rendering consistency with a version of multi-view attention, they cannot directly improve pose accuracy. The Zero123-XL [14] learns to capture pose better than the baseline [38] by training on the larger `Objaverse-XL`, but the pose is still *not* very accurate. Because of the RCN layer, Free3D shows no such pose errors and results in better images.

Method	OmniObject3D [73]					GSO [17]				
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PPLC↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PPLC↓
Zero-1-to-3 [38]	16.84	0.7813	0.1321	1.73	24.58	19.65	0.8501	0.0758	3.24	33.15
Zero123-XL [14]	17.11	0.7818	0.1291	1.51	21.33	20.43	0.8589	0.0706	3.23	28.03
SyncDreamer [39]	17.00	0.7941	0.1442	6.58	11.49	14.72	0.7835	0.1533	8.65	9.42
Consistent123 [69]*	17.13	0.7821	0.1255	1.55	18.02	20.11	0.8553	0.0716	3.24	20.08
Ours (Free3D)	**18.23**	**0.8090**	**0.0996**	**1.34**	**8.67**	**21.13**	**0.8686**	**0.0619**	**2.85**	**9.10**

Table 2. **Generalizable results on unseen datasets**, including `OmniObject3D` [73] and `GSO` [17], with 6,000 and 1,030 3D instances, respectively. Note that, although Zero123-XL [14] is trained on a larger dataset, and shows better generalizability, the proposed Free3D still significantly outperforms it with *precise* pose estimation for target views.



(a) Input View **(b)** Zero-1-to-3 [38] **(c)** Consistent123 [69] **(d)** Ours Free3D

Figure 6. **Qualitative new-view synthesis comparisons.** Zero-1-to-3 [38] generates diverse details (*e.g.* various ears and tails) for different views in one sampling. Consistent123 [69] improves it through the multi-view diffusion with the cross-view attention. However, it still requires training additional NeRF for 3D reconstruction. For more comparisons, see the supplemental videos for better visualisation.

4.3. Assessing Generalization

In Tab. 2 and Fig. 5, we validate the ability of Free3D to generalize to datasets not seen during training. This includes the `OmniObject3D` and the `GSO` datasets, with 6,000 scanned objects in 190 categories and 1,030 scanned objects in 17 categories, respectively. For this result, we directly test all trained models without any fine-tuning.

Quantitative comparison. In Tab. 2, Free3D, which uses the same `training set` as Zero-1-to-3 [38], outperforms the baseline and all state-of-the-art variants, including Zero123-XL [14], which is trained on a larger 3D dataset, and SyncDreamer [39], which utilizes a heavier 3D volumetric representation. SyncDreamer improves the baseline on `OmniObject3D`, which has small elevation views change, while achieving worse results on `GSO` with random elevation angles. Consistent123 [69] is also trained multiple views jointly. However, their relative improvement is limited. Table 2 shows that Free3D achieves very substantial improvements on all instantiations on both `OmniObject3D` and `GSO` datasets. This further indicates that ray conditioning can successfully improve the pose accuracy of the target view.

Qualitative comparison. A qualitative comparison is given in Fig. 5 (more in supplemental Figs. C.3 and C.4). Though Free3D is only trained on `objaverse` dataset, it works quite well in the open-set setting. Moreover, it shows significantly better results than all state-of-the-art models.

Results on real case. To further demonstrate the generalization of our Free3D to real images, following [36], we run Segment Anything [35] to extract the segments and show qualitative results in Fig. 1. Besides, we also show diverse results in Fig. 7 by sampling different noises.

4.4. Assessing 3D consistency

To assess 3D consistency, we render a 360° video with the fixed elevation angle $\theta := 0°$ and 50 azimuth angles uniformly sampled in $[0°, 360°]$. The quantitative results are reported in Tabs. 1 and 2 using the proposed PPLC score,

	Method	Objaverse [15]					GSO [17]				
		PSNR↑	SSIM↑	LPIPS↓	FID↓	PPLC↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PPLC↓
$\mathbb{A}$	Baseline Zero-1-to-3 [38]	19.65	0.8462	0.0938	1.52	22.10	19.65	0.8501	0.0758	3.24	33.15
$\mathbb{B}$	+ Input Ray Embeddings	20.21	0.8550	0.0858	1.26	16.63	20.49	0.8617	0.0677	3.01	22.08
$\mathbb{C}$	+ Multi-Scale Ray Emb.	20.56	0.8609	0.0797	1.30	15.94	20.50	0.8615	0.0667	2.98	20.09
$\mathbb{D}$	+ RCN	20.78	**0.8620**	0.0784	**1.21**	15.67	21.13	0.8686	0.0619	**2.85**	18.48
$\mathbb{E}$	+ Pseudo-3D attention	**20.81**	**0.8620**	**0.0781**	1.25	14.76	**21.20**	**0.8697**	**0.0617**	2.86	17.39
$\mathbb{F}$	$\mathbb{E}$ + noise sharing	—	—	—	—	**11.39**	—	—	—	—	**9.10**

Table 3. **Ablations** of Free3D design choices. Here, to reduce the computational cost and testing time, we only evaluate 2,000 instances in `Objaverse` dataset, instead of running the whole dataset with 7,729 for 50-frame 360° video rendering. Therefore, the PPLC score is slightly different from the values reported in Tab. 1.

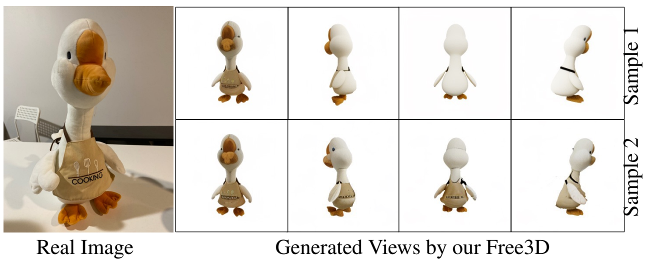


Figure 7. **Diverse NVS in real scenes.** Zoom in to see the details.

and the qualitative results are shown in Figs. 1 and 6. Consistent123 [69] still requires training additional NeRF to achieve 3D consistent video, while the directly rendered videos are still flickering. SyncDreamer [39] obviously improves view consistency by using an explicit 3D volume representation. However, it is trained only on a fixed elevation angle and can generate only 16 fixed frames for a video in its code. Interestingly, Free3D, by using *multi-view attention* and *multi-view noise sharing*, is nearly as effective (while being much cheaper). This is also partly due to ray conditioning, which can capture more precise poses of the target view and thus reduce ambiguity. The additional results are provided as videos in the supplemental materials.

4.5. Ablation Study

We run a number of ablations to analyse Free3D. Results are shown in Tab. 3 and discussed in detail next.

Our baseline configuration (denoted $\mathbb{A}$) is the same as Zero-1-to-3 [38], which is derived from the SD model [51] by replacing text conditioning with source image and target pose conditioning. Then, in $\mathbb{B}$ we extend this model with ray conditioning by concatenating the ray embeddings to the source image x^{src}. This alone improves the performance dramatically in both `Objaverse` and `GSO`. In $\mathbb{C}$ we test injecting ray conditioning into each level of the diffusion UNet ϵ_θ, but the generalizable performance remains similar to $\mathbb{B}$ in the unseen `GSO` dataset. In $\mathbb{D}$, we test instead the RCN layer, which results in further improvements on not only `Objaverse` dataset, but also on the new `GSO`. In $\mathbb{E}$, we add *multi-view attention* to exchange information between different frames. As expected, this does *not* improve the metrics that measure the quality of individual views, but it slightly improves the PPLC score, measuring consistency. In $\mathbb{F}$, we further add *multi-view noise sharing*, which significantly enhances the consistency between different views. while preserving the quality of single-view rendering. Rendered videos are provided in the supplement.

5. Conclusion

We have introduced Free3D, an open-set single-view NVS method with state-of-the-art performance on various categories, yet bypass the requirement of building on heavy 3D representation or training additional auxiliary 3D models. It is a simple approach that (i) obtains data prior from an off-the-shelf pre-trained 2D image generator, (ii) injects ray conditioning utilizing the new RCN layer to accurately code for the target pose, and (iii) combines that with multi-view attention and noise sharing to improve multi-view consistency. Experimental results show that Free3D significantly outperforms recent and concurrent state-of-the-art NVS models without incurring the cost of utilizing a 3D representation. We hope that Free3D will serve as a new strong baseline for single-image NVS and inspire future research in this area.

Acknowledgements. This research is supported by ERC-CoG UNION 101001212. Many thanks to Stanislaw Szymanowicz, Edgar Sucar, and Luke Melas-Kyriazi of VGG for insightful discussions and Ruining Li, Eldar Insafutdinov, and Yash Bhalgat of VGG for their helpful feedback. We would also like to thank the authors of Zero-1-to-3 [38] and Objaverse-XL [14] for their helpful discussions.

Ethics. We use the Objaverse [15], OmniObject3D [73], and GSO [17] following their terms and conditions. These datasets contain synthetic or scanned 3D objects, but, as far as we could determine, no personal data. For further details on ethics, data protection, and copyright please see https://www.robots.ox.ac.uk/~vedaldi/research/union/ethics.html.

References

[1] Edward H Adelson and John YA Wang. Single lens stereo with a plenoptic camera. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)*, 14(2):99–106, 1992. 3

[2] Edward H Adelson, James R Bergen, et al. The plenoptic function and the elements of early vision. *Computational models of visual processing*, 1(2):3–20, 1991. 3

[3] Titas Anciukevičius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Niloy J Mitra, and Paul Guerrero. Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12608–12618, 2023. 3

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5855–5864, 2021. 2

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5470–5479, 2022. 2

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 3, 4

[7] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023. 2, 4

[8] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 3

[9] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. 2

[10] Eric Ming Chen, Sidhanth Holalkere, Ruyu Yan, Kai Zhang, and Abe Davis. Ray conditioning: Trading photo-realism for photo-consistency in multi-view image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 4

[11] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, page 279–288, New York, NY, USA, 1993. Association for Computing Machinery. 2

[12] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. *arXiv preprint arXiv:2304.12294*, 2023. 3

[13] Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, 1996. 2

[14] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 5, 6, 7, 8, 3, 4

[15] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023. 1, 2, 3, 5, 8

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems (NeurIPS)*, 34:8780–8794, 2021. 1

[17] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 1, 2, 5, 6, 7, 8

[18] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *International Conference on Learning Representations (ICLR)*, 2016. 4

[19] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5501–5510, 2022. 2

[20] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[21] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2, 4

[22] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 5

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017. 5

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems (NeurIPS)*, 33:6840–6851, 2020. 1

[25] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2, 4

[26] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pages 1501–1510, 2017. 4

[27] Zixuan Huang, Stefan Stojanov, Anh Thai, Varun Jampani, and James M Rehg. Planes vs. chairs: Category-guided 3d shape learning without any 3d cues. In *European Conference on Computer Vision*, pages 727–744. Springer, 2022. 2

[28] Yifan Jiang, Hao Tang, Jen-Hao Rick Chang, Liangchen Song, Zhangyang Wang, and Liangliang Cao. Efficient-3dim: Learning a generalizable single-image novel-view synthesizer in one day. *arXiv preprint arXiv:2310.03015*, 2023. 4

[29] Angjoo Kanazawa, Shubham Tulsiani, Alexei A Efros, and Jitendra Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018. 2

[30] Yash Kant, Aliaksandr Siarohin, Michael Vasilkovsky, Riza Alp Guler, Jian Ren, Sergey Tulyakov, and Igor Gilitschenski. invs: Repurposing diffusion inpainters for novel view synthesis. *arXiv preprint arXiv:2310.16167*, 2023. 3, 5

[31] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy J Mitra. Holodiffusion: Training a 3d diffusion model using 2d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18423–18433, 2023. 3

[32] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4

[33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 2

[34] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. 2

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 7

[36] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 3, 7

[37] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Zexiang Xu, Hao Su, et al. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *arXiv preprint arXiv:2306.16928*, 2023. 2, 3, 4

[38] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023. 2, 3, 4, 5, 6, 7, 8, 1

[39] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Learning to generate multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 2, 3, 4, 5, 6, 7, 8, 1

[40] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 4

[41] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8446–8455, 2023. 2, 3

[42] Luke Melas-Kyriazi, Christian Rupprecht, and Andrea Vedaldi. Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12923–12932, 2023. 2

[43] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European conference on computer vision (ECCV)*, 2020. 2, 4, 1

[44] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[45] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition (CVPR)*, pages 3500–3509, 2017. 2, 3

[46] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 165–174, 2019. 2

[47] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2

[48] Senthil Purushwalkam and Nikhil Naik. Conrad: Image constrained radiance fields for 3d generation from a single image. *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 3

[49] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Sko-

rokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. *arXiv preprint arXiv:2306.17843*, 2023. 3
[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning (ICML)*, pages 8748–8763. PMLR, 2021. 4
[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 10684–10695, 2022. 2, 8, 1
[52] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 2
[53] Mehdi SM Sajjadi, Daniel Duckworth, Aravindh Mahendran, Sjoerd van Steenkiste, Filip Pavetic, Mario Lucic, Leonidas J Guibas, Klaus Greff, and Thomas Kipf. Object scene representation transformer. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:9512–9524, 2022. 3
[54] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6229–6238, 2022. 3
[55] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994*, 2023. 2, 3
[56] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022. 2, 1
[57] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 2, 3
[58] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2022. 2, 4
[59] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 32, 2019. 2
[60] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:19313–19325, 2021. 2, 4, 1
[61] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning (ICML)*, pages 2256–2265. PMLR, 2015. 2
[62] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8269–8279, 2022. 3
[63] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion: (0-)image-conditioned 3D generative models from 2D data. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 2, 3
[64] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023. 3
[65] Maxim Tatarchenko, Alexey Dosovitskiy, and Thomas Brox. Multi-view 3d models from single images with a convolutional network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 322–337. Springer, 2016. 3
[66] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16773–16783, 2023. 4
[67] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024. 3, 4
[68] Daniel Watson, William Chan, Ricardo Martin Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 4
[69] Haohan Weng, Tianyu Yang, Jianan Wang, Yu Li, Tong Zhang, CL Chen, and Lei Zhang. Consistent123: Improve consistency for one image to 3d object synthesis. *arXiv preprint arXiv:2310.08092*, 2023. 2, 3, 4, 5, 6, 7, 8
[70] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7467–7477, 2020. 3
[71] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu

Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1900–1910, 2023. 4

[72] Ruiqi Wu, Liangyu Chen, Tong Yang, Chunle Guo, Chongyi Li, and Xiangyu Zhang. Lamp: Learn a motion pattern for few-shot-based video generation. *arXiv preprint arXiv:2310.10769*, 2023. 2, 4

[73] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 803–814, 2023. 1, 2, 5, 6, 7, 8

[74] Yifeng Xiong, Haoyu Ma, Shanlin Sun, Kun Han, and Xiaohui Xie. Light field diffusion for single-view novel view synthesis. *arXiv preprint arXiv:2309.11525*, 2023. 3, 4

[75] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. *arXiv preprint arXiv:2403.14621*, 2024. 2, 3

[76] Jiayu Yang, Ziang Cheng, Yunfei Duan, Pan Ji, and Hongdong Li. Consistnet: Enforcing 3d consistency for multi-view images diffusion. *arXiv preprint arXiv:2310.10343*, 2023. 2, 3, 4, 5

[77] Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2024. 5

[78] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4578–4587, 2021. 2

[79] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[80] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018. 4, 5

[81] Chuanxia Zheng, Tung-Long Vuong, Jianfei Cai, and Dinh Phung. Movq: Modulating quantized vectors for high-fidelity image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:23412–23425, 2022. 4

[82] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12588–12597, 2023. 3